A Simple, Maintainable System for Identifying Subsets of a Modern Controlled Vocabulary

Erich B. Schulz M.B. B.S.*, James W. Barrett M.Sc., M.B. B.Chir.,
Alasdair R. Maclean B.Sc., F.R.C.A., Philip J. B. Brown, M.R.C.G.P.
NHS Centre for Coding and Classification, Loughborough, United Kingdom
*Now at National Centre for Classification in Health (Brisbane), QUT, Australia

The identification of subsets adds significant value to a large controlled vocabulary. However, they are potentially difficult to maintain as the hierarchy evolves. A simple scheme has been developed by the UK NHS Centre for Coding and Classification to enable the generation of subsets from a list of instructions. Capturing the intention of the subset creators as a list of instructions rather than as an explicit list of codes is hoped to facilitate maintenance. The scheme attempts to balance expressivity and simplicity.

Modern controlled vocabularies are large. There is a need to manage their size to reduce storage requirements, retrieval times, and duration of term searching. Time taken to enter coded data will be reduced by decreasing the size of picking lists and filtering out non-relevant codes during hierarchical navigation. This is facilitated by identification of subsets of codes to fulfil specific functions. Frequently, this consists of removing excessive specialist detail or excluding concepts outside of a single specialist domain. Some subsets do not follow specialty boundaries but are task oriented. Consistent with these goals, subsets have initially fallen into one of three types: generalist 'skirts' (figure 1a), specialist 'slivers' (figure 1b), and specific purpose 'speckled' sets (figure 1c).

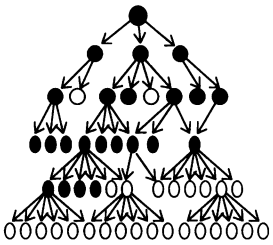

Figure 1a Skirt Subset

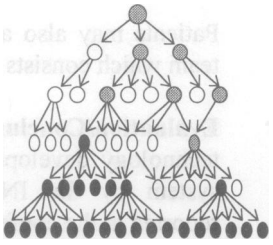

Figure 1b Sliver Subset

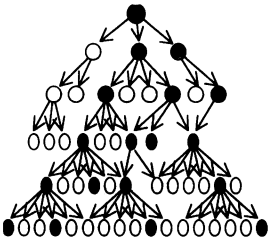

Figure 1c Speckle Subset

Key:

- ● Within Subset
- ◍ 'Grey' within Subset
- ○ Not in Subset

Early experience clearly demonstrates that choosing subsets is a subjective process and that iterative development will be required. During this phase hierarchies will change and new codes will be inserted in response to user feedback. Hopefully, expressing the subset authors' intention as a list of instructions rather than an explicit list of codes within the set will provide greater stability and easier maintenance.

Table 1 - Instruction types and order of execution

Instruction	Order	Description
INCUPDN	1	Include this node, all its ancestors and all its descendants
INCUP	2	Include this node and all its ancestors
INCDN	3	Include this node and all its descendants
EXDN	4	Remove this node and all its descendants
INC	5	Include this node only
EX	6	Remove this node only

Each instruction consists of a code (or node identifier) and one of the possible instructions shown in table 1. A generation script combines these instructions with the child-parent links within the hierarchy table to produce an explicit list of included codes. Order of execution becomes significant when removing codes from the set and this is also listed in table 1.

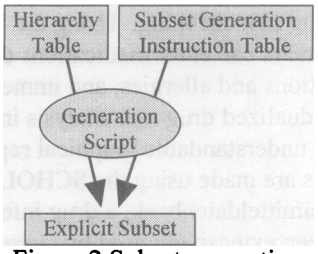

Figure 2 Subset generation process

After the initial set is generated additional nodes are added to enable all codes within the subset to be bound into a single hierarchy. This step is performed automatically by identifying all codes in the set without an immediate parent within the set and then marking all nodes along all paths from these nodes to the root node as 'grey'. The process of generating subsets suitable for implementation within clinical systems is summarised in Figure 2. Although subsets are a valuable resource, ideally users should be able to turn off the subset 'filter' if they fail to locate their desired code, even though this sacrifices the potential savings in size.

The simple scheme above is designed to facilitate continued enhancement of the subsets within an evolving Thesaurus, enabling their eventual incorporation into clinical systems.

1091—8280/97/$5.00